10/665,156

(19) Eur pälsches Patentamt
Eur pean Patent Offic
Offic europ´ n des brevets

(11) EP 1 004 968 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
31.05.2000 Bulletin 2000/22

(51) Int Cl.[7]: **G06F 17/30**

(21) Application number: **99309415.0**

(22) Date of filing: **25.11.1999**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **26.11.1998 JP 33627898**

(71) Applicant: **CANON KABUSHIKI KAISHA**
Tokyo (JP)

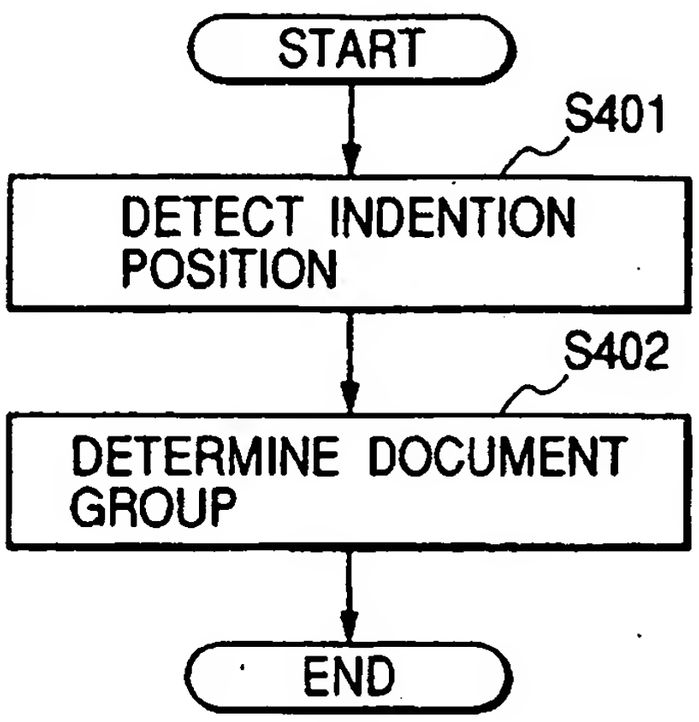
(72) Inventor: **Mizuno, Takafumi,**
c/o Canon Kabushiki Kaisha
Ohta-ku, Tokyo (JP)

(74) Representative:
Beresford, Keith Denis Lewis et al
BERESFORD & Co.
High Holborn
2-5 Warwick Court
London WC1R 5DJ (GB)

(54) Document type definition generating method and apparatus

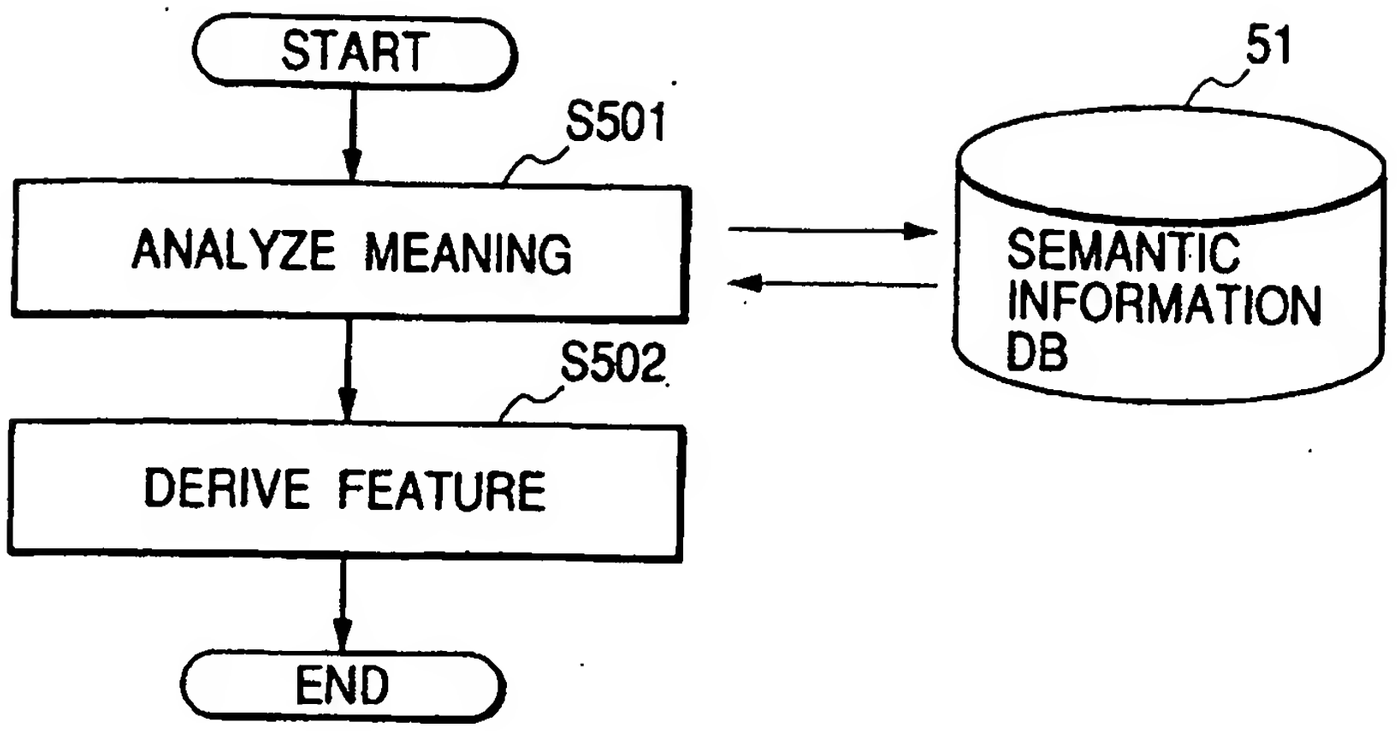
(57) There is disclosed a document type definition generating method comprising, in a structured document provided with a tag having an element name in each document element, judging a physical structure of each document element from indention, blank lines, and positional relation between tags, analyzing words and phrases in each document element, and judging a semantic structure of the document element based on words and phrases connection and word types. When the physical and semantic structures of document elements having tags different in element name are similar, the elements are regarded as being of the same type and one element name is excluded from a list for generating the document type definition. When the physical and semantic structures of document elements having tags with the same element name are different, the elements are regarded as being of the different types and one element name is changed. Furthermore, the words and phrases between a start tag and an end tag with the same title are analyzed, and the information to be included between the tags is obtained to generate the document type definition. Thereby, tag meaning is correctly treated, and the document type definition with tag r dundancy remov d therefrom is generated.

FIG. 4



EP 1 004 968 A2

FIG. 5


START
S501
ANALYZE MEANING
S502
DERIVE FEATURE
END
51
SEMANTIC INFORMATION DB

D scripti n

BACKGROUND OF THE INVENTION

Field of the Invention

[0001] The present invention relates to a computerized document processing executed by a personal computer, a word processor, and the like, particularly to a method and apparatus for generating the document type definition of a structured document, and a storage medium in which a program is stored.

Related Background Art

[0002] In recent years, the computerized documents prepared by a personal computer, a word processor, and the like have widely been used. The introduction of a structured document is advanced in which the computerized document is consistently treated and the elements constituting the document are provided with semantic information. In this structured document, each document element is held between front and back tags including element names (tag names), and in many cases description is performed for each document type in accordance with the document type definition of defining a place, order, frequency and the like in which the element appears.

[0003] On the other hand, the structured document can be described without preparing the document type definition. However, when the documents prepared by a plurality of users are integrated to form one document, and if the individual users use the tags having arbitrary titles, there is a possibility of attaching different tag names to the same element, or conversely attaching the same tag name to different elements.

[0004] In this case, there arise problems that the semantic information attached to the tag cannot correctly be handled, and that redundancy is generated with respect to the tag.

SUMMARY OF THE INVENTION

[0005] An objective of the present invention is to provide a method and apparatus for generating document type definition from a structured document provided with tags, and a storage medium which stores the program.

[0006] Another objective of the present invention is to provide a document type definition generating method and apparatus which can correctly treat semantic information given to tags, and a storage medium which stores the program.

[0007] Further objectiv of the present invention is to provide a document typ d finition g nerating method and apparatus which can g nerate docum nt type definition with redundancy to tags removed th r from, and a storage medium which stores the program.

[0008] According to one aspect, the present invention which achieves these objectives relates to a document processing method comprising: in a structured document provided with a tag having an element name in each document element, a physical structure judging step of judging a physical structure of each document element; a semantic structure judging step of judging a semantic structure of the document element; and a document type definition generating step of generating document type definition to define appearance state of the document element in the structured document based on judgment results of the physical structure judging step and the semantic structure judging step.

[0009] According to another aspect, the present invention which achieves these objectives relates to a document processing apparatus comprising: in a structured document provided with a tag having an element name in each document element, physical structure judging means for judging a physical structure of each document element; semantic structure judging means for judging a semantic structure of the document element; and document type definition generating means for generating document type definition to define appearance state of the document element in the structured document based on judgment results of the physical structure judging means and the semantic structure judging means.

[0010] According to still another aspect, the present invention which achieves these objectives relates to a computer-readable storage medium storing a document type definition generating program for controlling a computer to perform document type definition generation, the program comprising codes for causing the computer to perform, in a structured document provided with a tag having an element name in each document element, a physical structure judging step of judging a physical structure of each document element, a semantic structure judging step of judging a semantic structure of the document element, and a document type definition generating step of generating document type definition to define appearance state of the document element in the structured document based on judgment results of the physical structure judging step and the semantic structure judging step.

[0011] Other objectives and advantages besides those discussed above shall be apparent to those skilled in the art from the description of a preferred embodiment of the invention which follows. In the description, reference is made to accompanying drawings, which form a part thereof, and which illustrate an example of the invention. Such example, however, is not exhaustive of th various embodiments of the inv ntion, and th r fore reference is made to the claims which follow the description for d termining the scope of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] Fig. 1 is a block diagram of a document type definition generating apparatus.

[0013] Fig. 2 is a flowchart showing the procedure of a document type definition generation processing.
[0014] Figs..3A and 3B are diagrams showing examples of structured document data.
[0015] Fig. 4 is a flowchart showing the processing procedure of physical structure analysis.
[0016] Fig. 5 is a flowchart showing the processing procedure of semantic structure analysis.
[0017] Fig. 6 is a flowchart showing the processing procedure of removing tag redundancy.
[0018] Fig. 7 is a diagram showing one example of document type definition.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0019] A preferred embodiment of the present invention will be described hereinafter with reference to the accompanying drawings.

<First Embodiment>

[0020] Fig. 1 is a block diagram of a document type definition generating apparatus according to the present invention.
[0021] In Fig. 1, an input unit 101 is constituted of a keyboard, a pointing apparatus, and the like, and is used for a user to input data or commands. An external memory unit 102 is constituted of a storage apparatus using media such as a hard disk to store structured document data as a processing object, data of semantic information database (DB) described later, generated document type definition, and the like. A display unit 103 is constituted of CRT, a liquid crystal display, and the like to display the structured document data, the generated document type definition, and the like.
[0022] A CPU 104 performs control of each component of the apparatus, reads and executes a program, and realizes various processings. A ROM 105 stores fixed data and program. A control program for realizing a processing procedure as described later with reference to the flowcharts of Fig. 2 to 6 may be stored in the ROM 105, or read from the external memory unit 102. A RAM 106 presents an operation area necessary for the processing of the apparatus. A bus 107 connects the apparatus components.
[0023] Fig. 2 is a flowchart showing the procedure of a document type definition generation processing according to the present invention.
[0024] First, the structured document is inputted in step S201. This is executed by reading th structured document from the external m mory unit 102. One x-ample of the structured document given her in is shown in Fig. 3A. For xample, a first lin "<Titl >" indicat s a start tag, "</Title>" indicates an end tag, and "TV SET OPERATING INSTRUCTIONS" h ld between these tags is a document element indicating a tag content. Moreov r, "Title" is an element name (tag name). Furthermore, the attribute and value of the element can be described in the tag.
[0025] In the next step S202, each tag position is detected from the structured document, and a tag number is attached in order from the top "<Title>".
[0026] Subsequently, in step S203, the physical structure in the document is detected. For example, in Fig. 3B, as diagrammatically represented in "<Para>" indicating a paragraph, a feature that a sentence group starting with an indention is regarded as the paragraph is detected. The processing procedure for detecting such physical structure is shown in the flowchart of Fig. 4.
[0027] First, in step S401, a line in which indention is performed is found in the document, and in the next step S402 the sentence group following the line is detected. In this case, the line in which the indention is performed to the line in which the next indention is performed, or to the line right before a blank line can be set to the sentence group. In this case, the indention (double indention) performed in quotation in which the quotation is represented by performing the indention, and blank lines described by constantly skipping one or more lines are excluded as structures meaningless for the detection of the physical structure from the entire document pattern to perform the processing in the step S402.
[0028] Turning back to Fig. 2, in the next step S204, the semantic structure of the inputted structured document is detected. As one example, in Fig. 3A, the contents of tags "<Section>" have forms in which "1.", "2.", "3." are attached to top positions. Here, the content of tag "<Section>" can semantically be presumed to have "numeral." on its top. One example of processing procedure for detecting the semantic structure is shown in the flowchart of Fig. 5.
[0029] First, in step S501, communication is performed with a semantic information database (DB) 51 with respect to all words and codes in the document to provide the connection between words in the document and the types of words and codes. In the next step S502, the semantic structure found in each document element is detected based on this result.
[0030] Returning to Fig. 2, in the next step S205, a first appearing tag is regarded as the tag to be processed, and it is judged in step S206 whether or not the processing of the tag is all completed.
[0031] When the tag processing is not completed, the process shifts to step S207, in which the tag as the present processing object, and the information on the physical and semantic structures detected in the steps S203 and 204 are unified. H r , the unifying m ans that when physical and s mantic features are present in the line related with the tag used as the pr sent processing obj ct, the tag and the information are connected. Subsequently, in step S208, the process is mov d to the n xt appearing tag, thereby returning to the step S206.
[0032] On the other hand, when it is judged in the step S206 that the tag processing is all completed, the proc-

ess shifts to step S209, in which similarity is obtained between the tags having different titles. When the similarity is equal to or more than a predetermined threshold value, the tags are regarded as the same tag, and one of the tags is prevented from appearing on the document type definition to be generated. The processing procedure for obtaining this similarity to determine whether or not the tags have the same content is shown in the flowchart of Fig. 6.

[0033] First, the similarity of tags A, B having different titles is calculated in step S601. This calculating method comprises setting the similarity of the physical structure to 1 when the physical structures agree with each other. When the physical structures do not completely agree with each other, but partially agree with each other, the similarity of the physical structure is set to a value less than 1 which corresponds to the agreed proportion. The similar concept is applied to the semantic structure, and the similarity of the semantic structure is obtained. The dividing of the sum of the similarity of the physical structure and the similarity of the semantic structure by 2 results in a general similarity d_{AB} of A and B.

[0034] In the next step S602, the similarity d_{AB} obtained in the step S601 is compared with the predetermined threshold value δ. When the similarity d_{AB} is less than δ, the process jumps to step S604 for trial of the next combination.

[0035] When the similarity d_{AB} is equal to or more than the threshold value δ, the process shifts to step S603, in which the tag B is regarded as being of the same type as the tag A, the tag B is finally struck off a list for generating the document type definition, and redundancy is removed.

[0036] When the processing of the step S603 is completed, the process advances to step S604, in which it is judged whether or not the trial of combination of all tags is made. When the combination of all tags is not tried, the process returns to the step S601. When the combination of all tags is tried, the subroutine processing is ended to return to the main routine of Fig. 2.

[0037] Moreover, in the step S209, in addition to the above-described processing of Fig. 6, the physical structure and semantic structure of the document elements having the same title are compared. When the structures are different, the title of one of the tags is changed. For this purpose, the similarity is obtained between tags Aa and Ab having the same tag name in the same manner as described above. When similarity value d_{AaAb} is less than the threshold value, the title of the tag Ab is changed. This threshold value may be differ nt from the above-described value.

[0038] In step S210 of Fig. 2, the sent nce word between the start tag and th end tag which have the same title is analyz d to obtain th information to be includ d in the tags. This analysis result is us d to gen rate the document type definition in the n xt step S211.

[0039] Fig. 7 is a diagram showing one example of the generated document type definition, and the document type definition generated from the structured document data shown in Fig. 3A is shown as document type "manual".

[0040] Here, in Fig. 3A, the content of tag <Sect> agrees in physical structure with the content of tag <Section>, and the tags are the same in semantic structure in that they have the form of "numeral.". Therefore, it is determined in the step S209 that the tag <Sect> has the same content as that of the tag <Section>. As a result, the generated document type definition does not use <Sect>, and in <Body>, Section+, that is, tag <Section> repeatedly appears.

<Second Embodiment>

[0041] In the above-described first embodiment, the physical and semantic structures in the document are judged based on the sentence (portions other than tags), but the present invention is not limited to this.

[0042] For example, the physical information such as the relative positional relation between the tags and the inclusive relation of the tags is detected as the physical structure, or the meaning represented by the tag name or attribute is detected as the semantic structure, so that these structures may be used as the objects to obtain the similarity.

[0043] According to the embodiments described above, since the physical and semantic structures of the document element surrounded with the tags are judged, and the document type definition of the structured document provided with the tags is generated, the semantic information given to the tags can correctly be treated.

[0044] Furthermore, the redundancy to the tags having the same content can be removed, and the document type definition can be generated in which there are no tags being the same in title and different in meaning.

[0045] Additionally, the present invention may be applied to a computer system constituted of a plurality of apparatuses (e.g., host computer, interface apparatus, reader, printer, and the like), or to a device constituted of one apparatus (e.g., word processor, copying machine, facsimile device, and the like).

[0046] Moreover, it goes without saying that the objective of the present invention can be achieved by supplying a storage medium storing the program code of software to realize the function of the above-described embodiment to the system or the device, and reading and executing the program code stored in the storage medium by the computer (or CPU or MPU) of the system or the d vice.

[0047] In this case, th program code its lf read from th storage medium realizes the function of the above-described embodiment, and the storag medium in which th program cod is r corded constitut s th pr s nt invention.

[0048] As the storag medium in which the program code, and tables and other variable data are stored, for xample, a floppy disk (FD), a hard disk, an optical disk,

an optomagnetic disk, CD-ROM, CD-R, a magnetic tape, a nonvolatile memory card (IC memory card), ROM, and the like can be used.

[0049] Moreover, the function of the above-described embodiment is realized by executing the program code read by the computer, but it goes without saying that the present invention also includes a case in which an operating system (OS) operating on the computer performs a part or the whole of an actual processing based on the instruction of the program code and the function of the above-described embodiment is realized by the processing.

[0050] Although the present invention has been described in its preferred from with a certain degree of particularity, many apparently widely different embodiments of the invention can be made without departing from the spirit and the scope thereof. It is to be understood that the invention is not limited to the specific embodiments thereof except as defined in the appended claims.

[0051] Additional aspects and embodiments of the invention are envisaged in which the document type definition generation step is based on only a judging step of judging physical structure. There is disclosed an alternative system in which the document type definition generation step is based on only a judging step of judging a semantic structure.

[0052] Further, the computer program for implementing the invention can be obtained in electronic form for example by downloading the code over a network such as the Internet. Thus in accordance with another aspect of the present invention there is provided an electrical signal carrying processor implementable instructions for controlling a processor to carry out the method as hereinbefore described.

Claims

1. A document type definition generating method, comprising, in a structured document provided with a tag having an element name in each document element:

   a physical structure judging step of judging a physical structure of each document element;
   a semantic structure judging step of judging a semantic structure of said each document element; and
   a document type definition gen rating step of generating docum nt type definition to defin appearance state of the document element in said structured document based on judgm nt r sults of said physical structur judging step and said s mantic structure judging step.

2. The document type definition generating method according to claim 1, wherein said physical structure judging step comprises judging the physical structure of the document element based on an indention or a blank line.

3. The document type definition generating method according to claim 2, wherein when the physical structure of the document element is judged based on said indention, the judging is performed by excluding the indention which represents quotation.

4. The document type definition generating method according to claim 2, wherein when the physical structure of the document element is judged based on said blank line, the judging is performed by excluding the blank line from a document in which description is made by constantly placing every predetermined number of blank lines.

5. The document type definition generating method according to claim 1, wherein said physical structure judging step comprises judging the physical structure of the document element based on a positional relation of the tags surrounding the document element.

6. The document type definition generating method according to claim 1, wherein said semantic structure judging step comprises referring to a semantic information database to judge the semantic structure of the document element based on words and phrases connection in a document and word types.

7. The document type definition generating method according to claim 1, wherein said semantic structure judging step comprises judging the semantic structure of the document element based on a meaning represented by the tags surrounding the document element.

8. The document type definition generating method according to claim 1, wherein said document type definition generating step comprises a redundancy removing step of, when the physical structure and the semantic structure of a plurality of document elements having the tags different in element name are similar, regarding the document elements as being of the same type and excluding one element name from a document type definition generating object based on the judgment results of said physical structure judging st p and said semantic structure judging step.

9. The document type definition generating method according to claim 8, wh rein said r dundancy removing st p comprises obtaining similarity degr es concerning agreement degr es of the physical structure and the semantic structure betwe n the document elements having the tags different in el-

ement name, and regarding the document elements as being of the same type when a general similarity value calculated from the similarity degrees is equal to or more than a predetermined threshold value.

10. The document type definition generating method according to claim 1, wherein said document type definition generating step comprises a title changing step of, when the physical structure and the semantic structure of a plurality of document elements having the tags with the same element name are different, regarding the document elements as being of different types and changing one element name based on the judgment results of said physical structure judging step and said semantic structure judging step.

11. The document type definition generating method according to claim 1, wherein said document type definition generating step comprises analyzing words and phrases present between a start tag and an end tag having the same title, obtaining information to be included between the tags, and generating the document type definition based on the information.

12. A document type definition generating apparatus comprising: in a structured document provided with a tag having an element name in each document element,

physical structure judging means for judging a physical structure of said each document element;
semantic structure judging means for judging a semantic structure of said each document element; and
document type definition generating means for generating document type definition to define appearance state of the document element in said structured document based on judgment results of said physical structure judging means and said semantic structure judging means.

13. The document type definition generating apparatus according to claim 12, wherein said physical structure judging means judges the physical structure of the document element based on an indention or a blank line.

14. The document type definition generating apparatus according to claim 13, wh rein said physical structure judging means judg s the physical structure of th docum nt l m nt bas d on said indention by excluding the indention which repres nts quotation.

15. Th document type definition generating apparatus according to claim 13, wh r in said physical structure judging means judges th physical structure of the document element based on said blank lines by excluding the blank lines from a document in which description is made by constantly placing every predetermined number of blank lines.

16. The document type definition generating apparatus according to claim 12, wherein said physical structure judging means judges the physical structure of the document element based on a positional relation of the tags surrounding the document element.

17. The document type definition generating apparatus according to claim 12, wherein said semantic structure judging means refers to a semantic information database to judge the semantic structure of the document element based on words and phrases connection in a document and word types.

18. The document type definition generating apparatus according to claim 12, wherein said semantic structure judging means judges the semantic structur of the document element based on a meaning represented by the tags surrounding the document element.

19. The document type definition generating apparatus according to claim 12, wherein said document type definition generating means comprises redundancy removing means for, when the physical structure and the semantic structure of a plurality of document elements having the tags different in element name are similar, regarding the document elements as being of the same type and excluding one element name from a document type definition generating object based on the judgment results of said physical structure judging means and said semantic structure judging means.

20. The document type definition generating apparatus according to claim 19, wherein said redundancy removing means obtains similarity degrees concerning agreement degrees of the physical structure and the semantic structure between the document elements having the tags different in element name, and regards the document elements as being of the same type when a general similarity value calculated from the similarity degrees is equal to or mor than a predetermined threshold value.

21. The document type definition generating apparatus according to claim 12, wh rein said document type definition gen rating means comprises title changing means for, when the physical structure and the semantic structure of a plurality of docum nt elem nts having the tags with the sam el ment name are different, regarding the document lements as being of different types and changing one element

name based on the judgment results of said physical structure judging means and said semantic structure judging means.

22. The document type definition generating apparatus according to claim 12, wherein said document type definition generating means analyzes words and phrases present between a start tag and an end tag having the same title, obtains information to be included between the tags, and generates the document type definition based on the information.

23. A computer-readable storage medium storing a document type definition generating program for controlling a computer to perform document type definition generation, said program comprising codes for causing the computer to perform:

in a structured document provided with a tag having an element name in each document element, a physical structure judging step of judging a physical structure of each document element:
a semantic structure judging step of judging a semantic structure of said each document element; and
a document type definition generating step of generating document type definition to define appearance state of the document element in said structured document based on judgment results of said physical structure judging step and said semantic structure judging step.

24. A method of removing redundancy in tags associated with elements of a document comprising comparing tags to obtain a similarity value, comparing the similarity value with a threshold to detect redundancy, and cancelling one of the tags if redundancy is detected.

25. A method of removing redundancy in tags associated with elements of a document comprising analysing elements having the same tags and cancelling one of the tags if the elements are identical.

26. A document type definition generating method, comprising, in a structured document provided with a tag having an element name in each document element:

a physical structur judging step of judging a physical structure of each document element;
a document typ d finition gen rating step of gen rating document type definition to defin appearance state of the document element in said structur d docum nt based on judgment results of said physical structure judging step.

27. A document type definition generating method, comprising, in a structured document provided with a tag having an element name in each document element:

a semantic structure judging step of judging a semantic structure of said each document element; and
a document type definition generating step of generating document type definition to define appearance state of the document element in said structured document based on judgment results of said semantic structure judging step.

28. An electrical signal carrying processor implementable instructions for controlling a processor to carry out the method of any one of claims 1 to 11.

FIG. 1

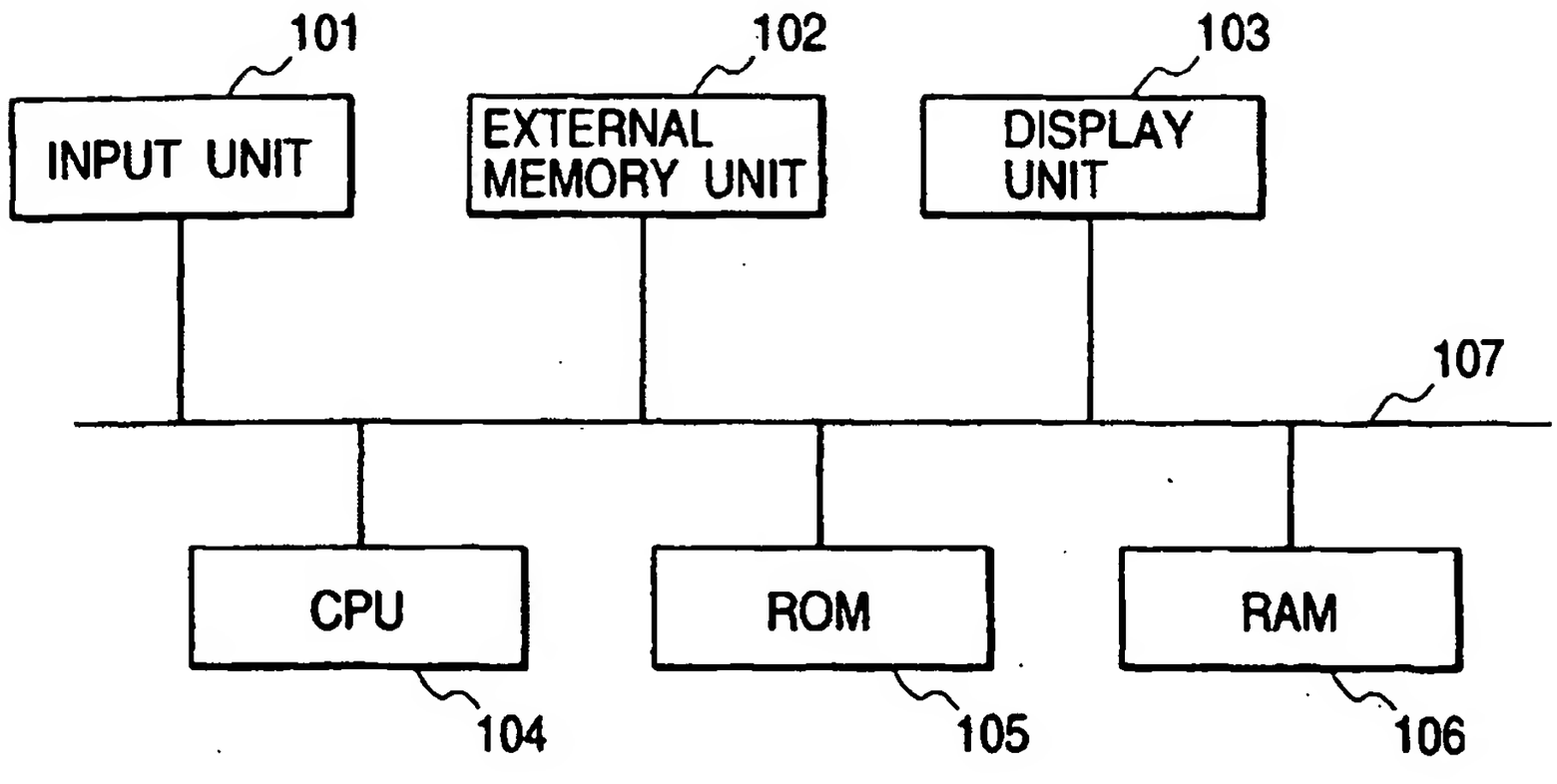

101
INPUT UNIT
102
EXTERNAL MEMORY UNIT
103
DISPLAY UNIT
107
CPU
ROM
RAM
104
105
106

FIG. 2

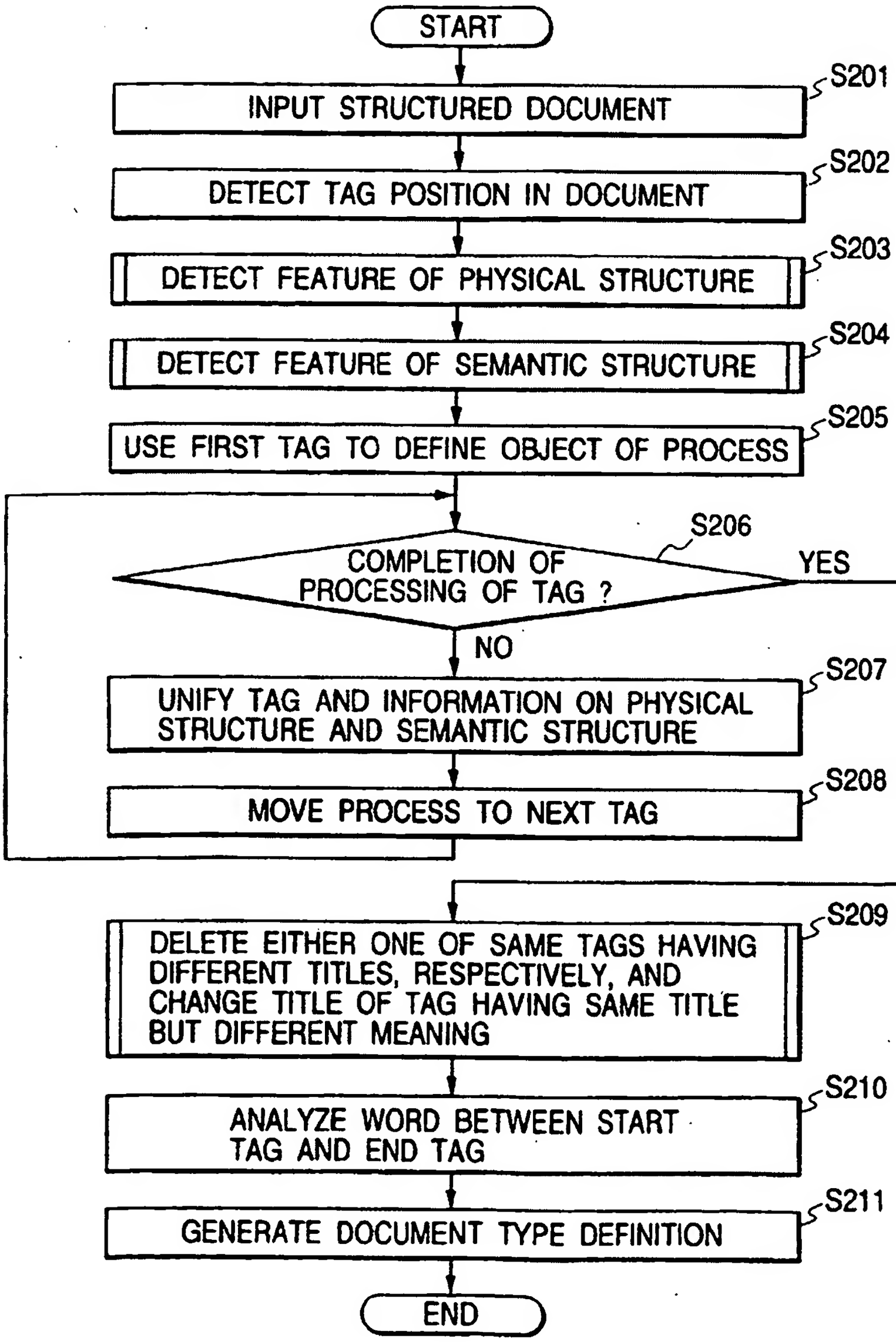

START
INPUT STRUCTURED DOCUMENT
S201
DETECT TAG POSITION IN DOCUMENT
S202
DETECT FEATURE OF PHYSICAL STRUCTURE
S203
DETECT FEATURE OF SEMANTIC STRUCTURE
S204
USE FIRST TAG TO DEFINE OBJECT OF PROCESS
S205
COMPLETION OF PROCESSING OF TAG ?
S206
YES
NO
UNIFY TAG AND INFORMATION ON PHYSICAL STRUCTURE AND SEMANTIC STRUCTURE
S207
MOVE PROCESS TO NEXT TAG
S208
DELETE EITHER ONE OF SAME TAGS HAVING DIFFERENT TITLES, RESPECTIVELY, AND CHANGE TITLE OF TAG HAVING SAME TITLE BUT DIFFERENT MEANING
S209
ANALYZE WORD BETWEEN START TAG AND END TAG
S210
GENERATE DOCUMENT TYPE DEFINITION
S211
END

FIG. 3A

```
<Title>TV SET OPERATING INSTRUCTIONS</Title>
<Date>19998.2.1</Date>
<Author>TARO YAMADA</Author>
<Body>
   <Section>1. PLUG IN</Section>
   <Section>2. TURN ON POWER</Section>
   <Section>3. TUNE IN</Section>
   <Sect>4. CONTROL VOLUME</Sect>
</Body>
```

FIG. 3B

```
<Para>
     ----------
   ----------
   ----------
</Para>

<Para>
     ----------
   ----------
</Para>
```

FIG. 4

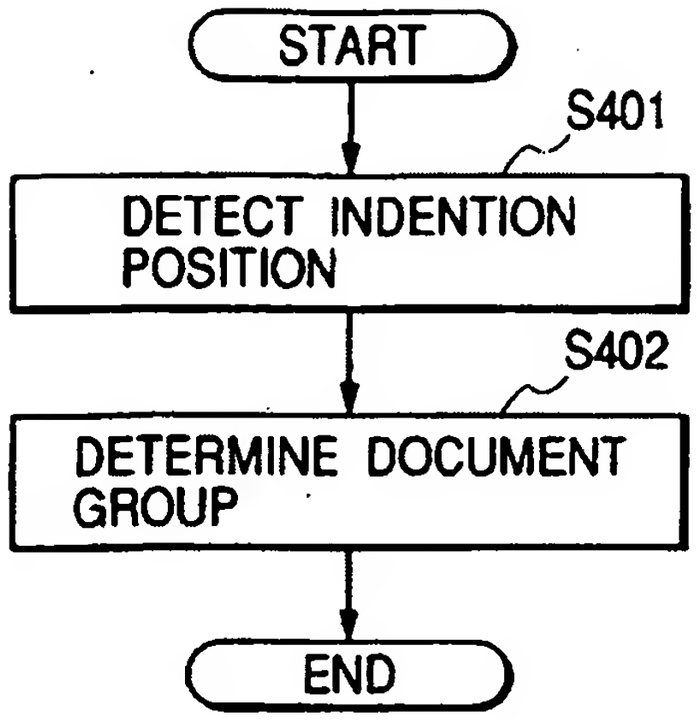

START
S401
DETECT INDENTION POSITION
S402
DETERMINE DOCUMENT GROUP
END

FIG. 5

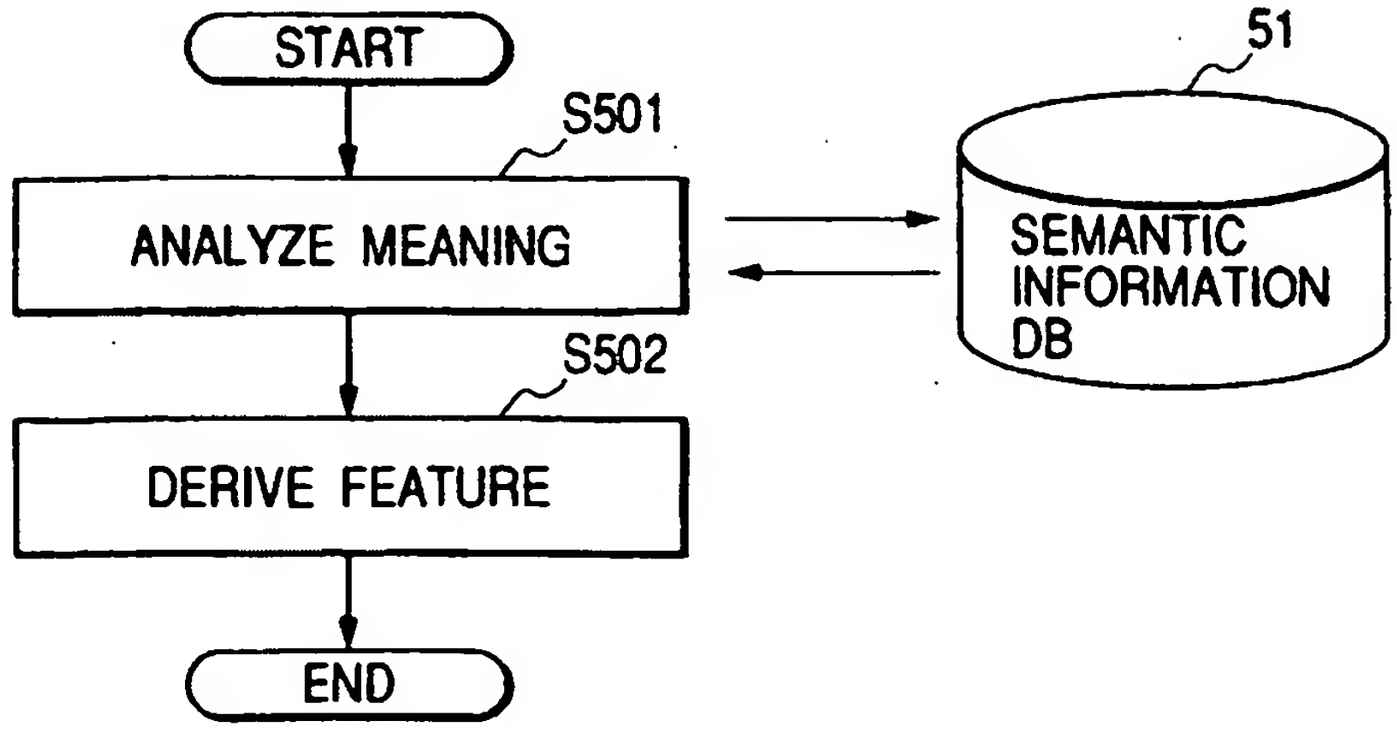

START
S501
ANALYZE MEANING
S502
DERIVE FEATURE
END
51
SEMANTIC INFORMATION DB

FIG. 6

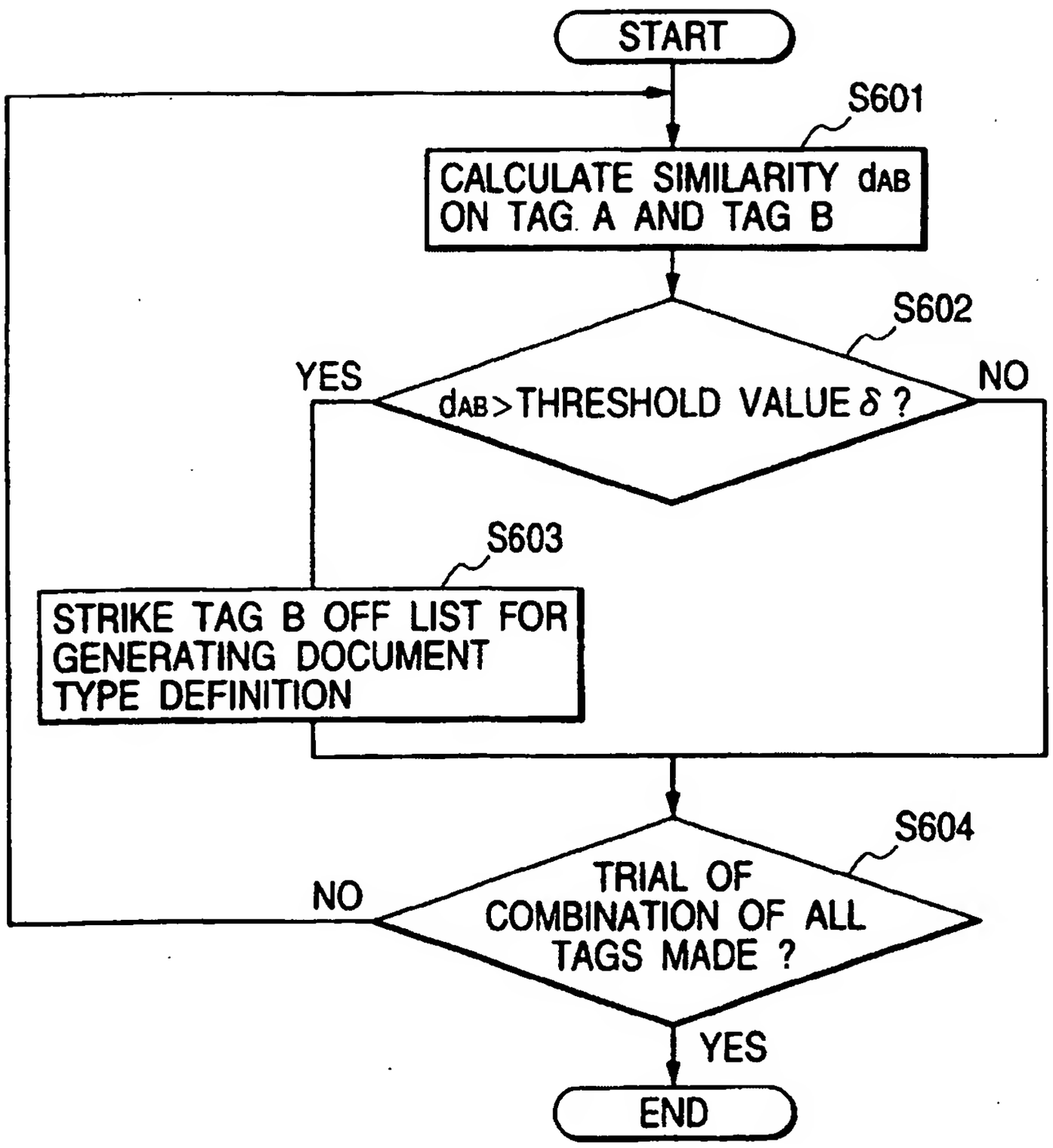

START
S601
CALCULATE SIMILARITY dAB ON TAG A AND TAG B
S602
dAB>THRESHOLD VALUE δ ?
YES
NO
S603
STRIKE TAG B OFF LIST FOR GENERATING DOCUMENT TYPE DEFINITION
S604
TRIAL OF COMBINATION OF ALL TAGS MADE ?
NO
YES
END

FIG. 7

```
<!DOCTYPE manual [
<!ELEMENT manual (Title, Date, Author, Body)>
<!ELEMENT Title (#PCDATA)>
<!ELEMENT Date (#PCDATA)>
<!ELEMENT Author (#PCDATA)>
<!ELEMENT Body (Section+)>
<!ELEMENT Section (#PCDATA)>
]>
```